Data Exploration

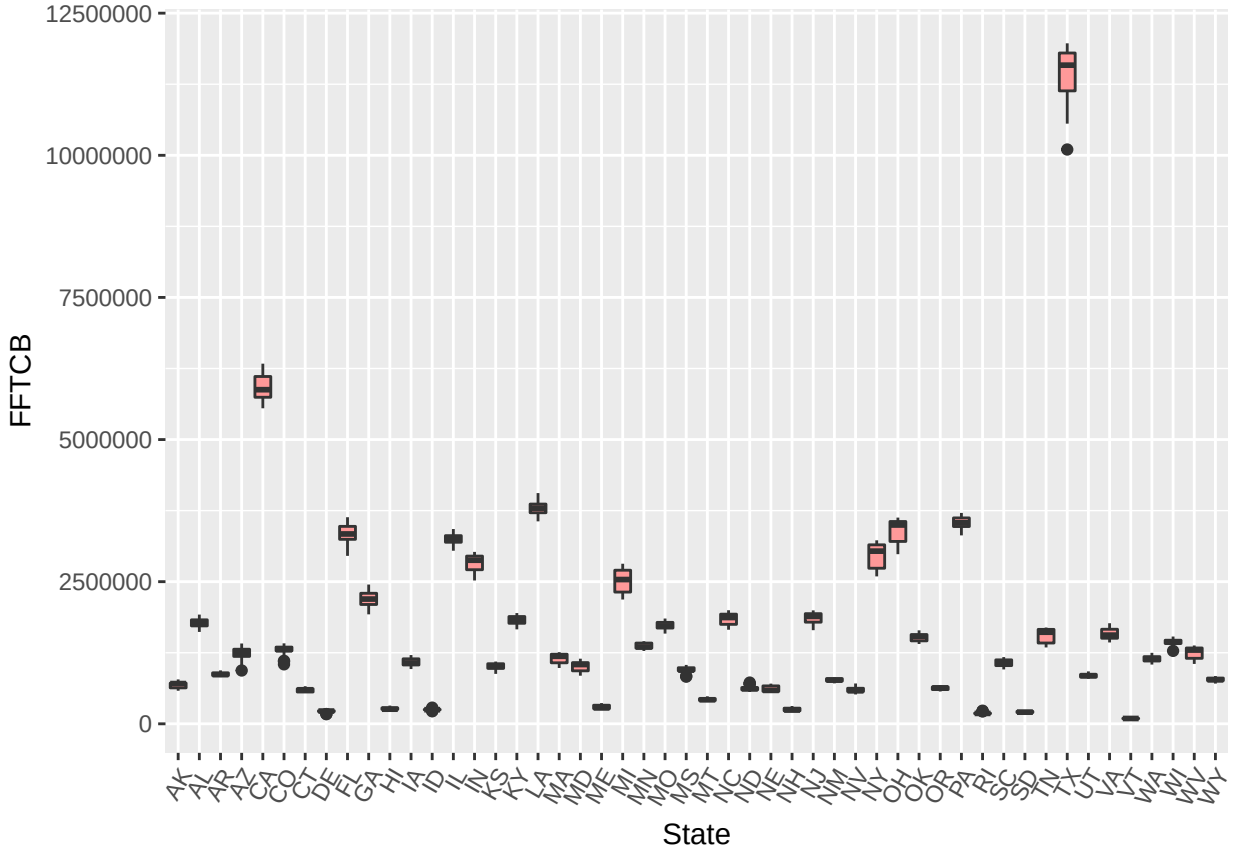
Eduardo Gomez

4/13/2018

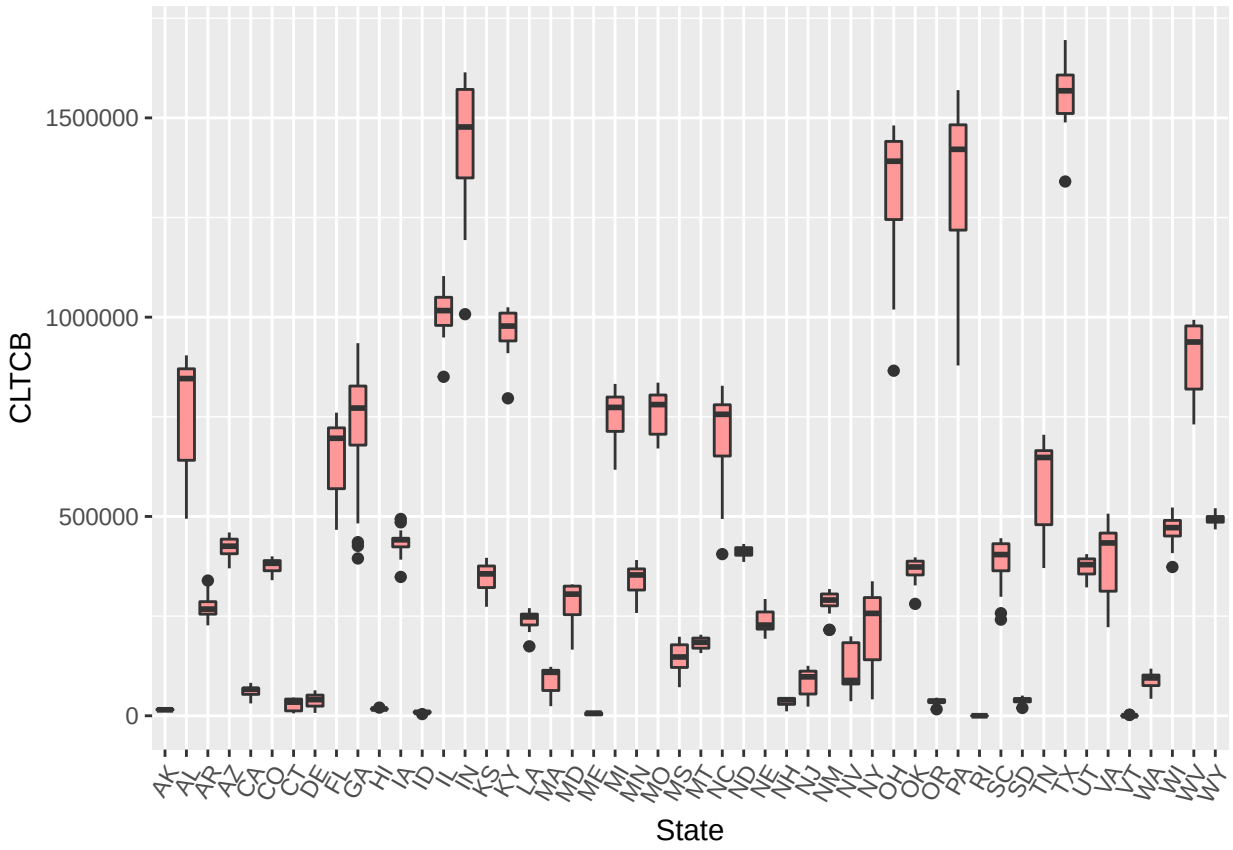
```
#using the states data
df <- read.csv("~/Energy_Analysis/data/usa_states_energy.csv")
```

Boxplot of how each states' growth is incomparision to each other.

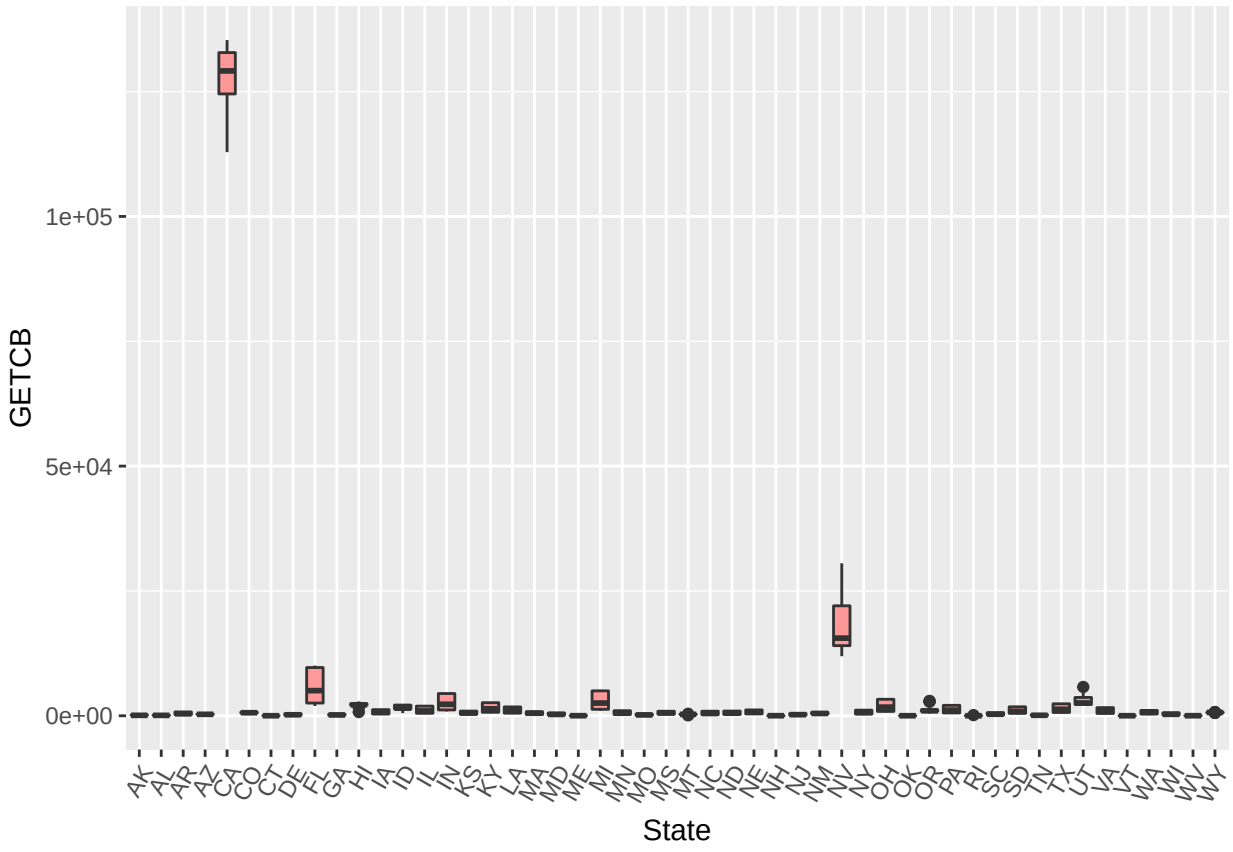
```
ggplot(data = df, aes(x = State, y = FFTCB)) + # fossil fuels
        geom_boxplot(fill = "#FF9999") +
    theme(axis.text.x = element_text(angle = 60, hjust = 1))
```




```
ggplot(data = df, aes(x = State, y = CLTCB)) + # coal
        geom_boxplot(fill = "#FF9999") +
    theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

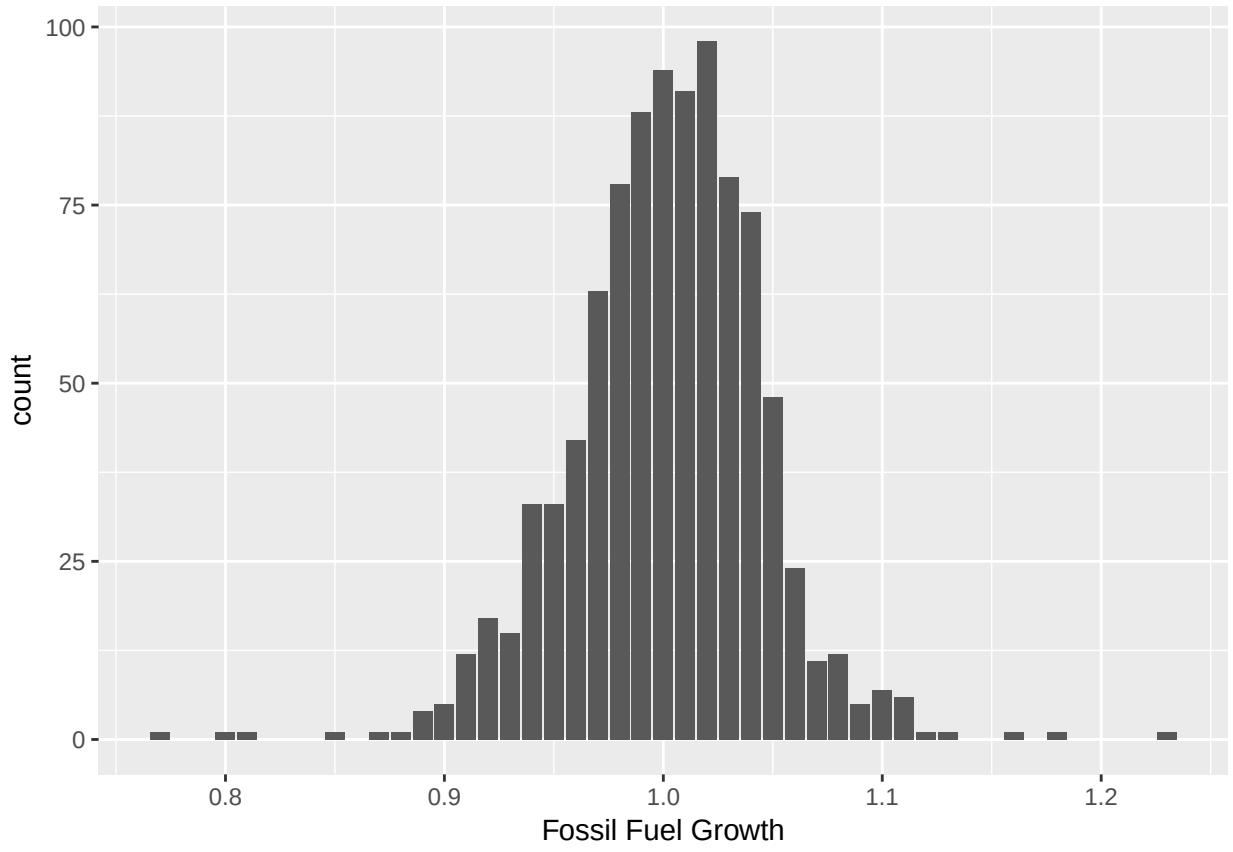



```
ggplot(data = df, aes(x = State, y = GETCB)) + #geothermal; california is thriving; not v popular
        geom_boxplot(fill = "#FF9999") +
    theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

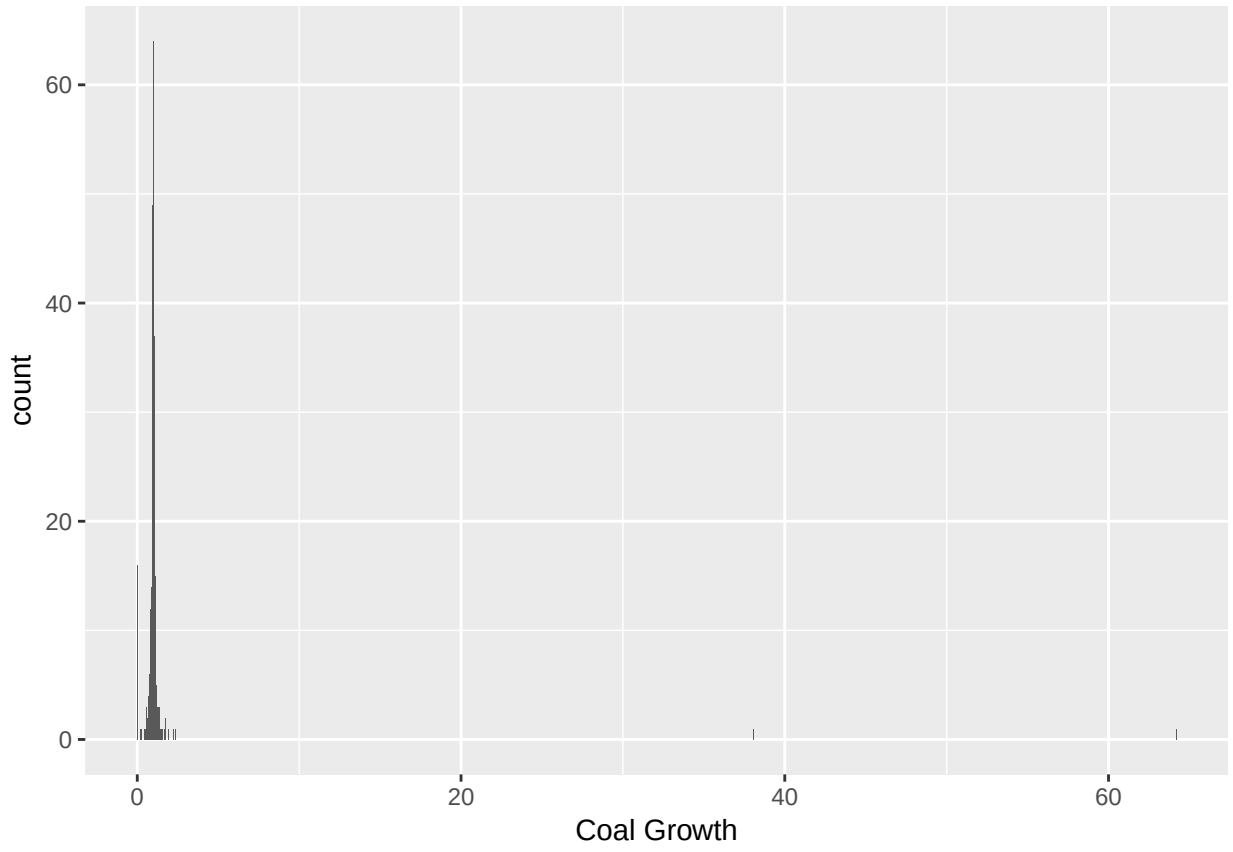


Distribution of covariates

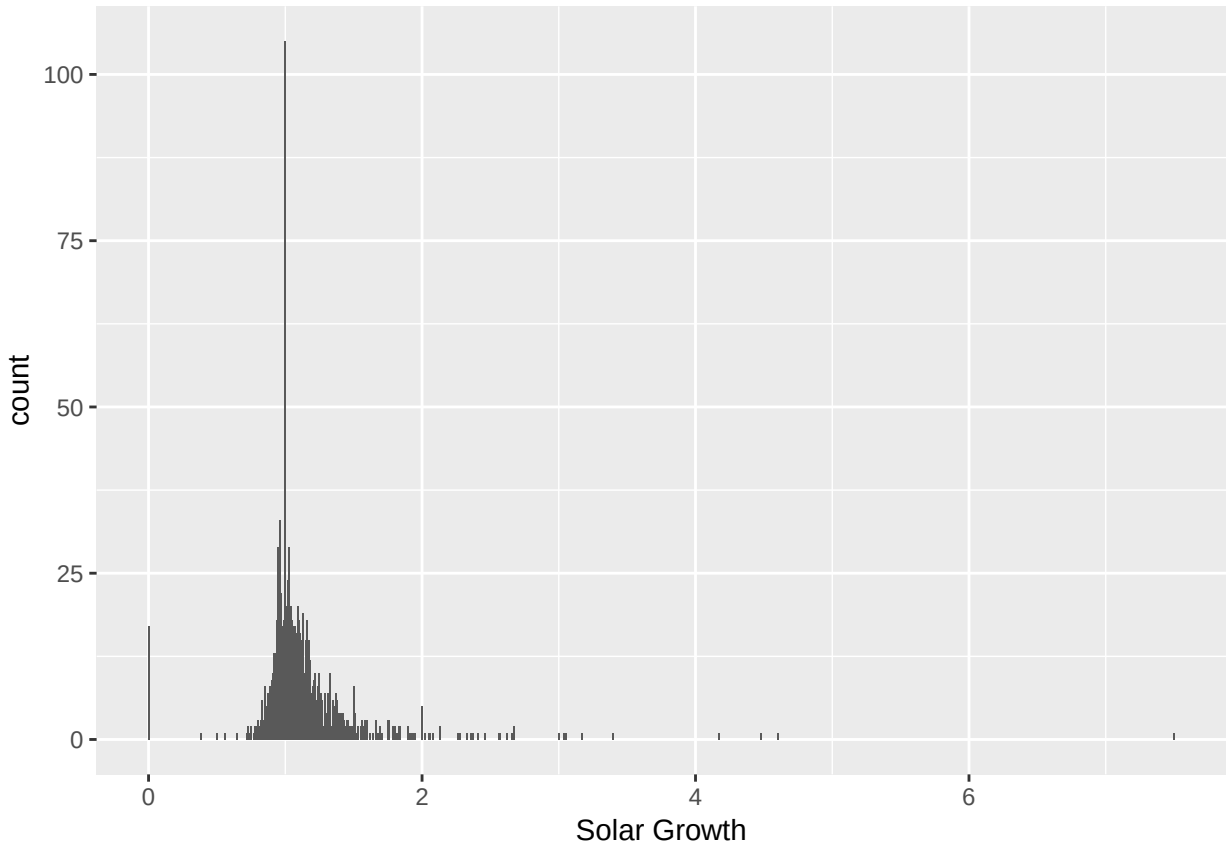
```
ggplot(df, aes(x=fossil_fuel.growth)) +
  geom_bar() + #normal
  xlab("Fossil Fuel Growth")
```



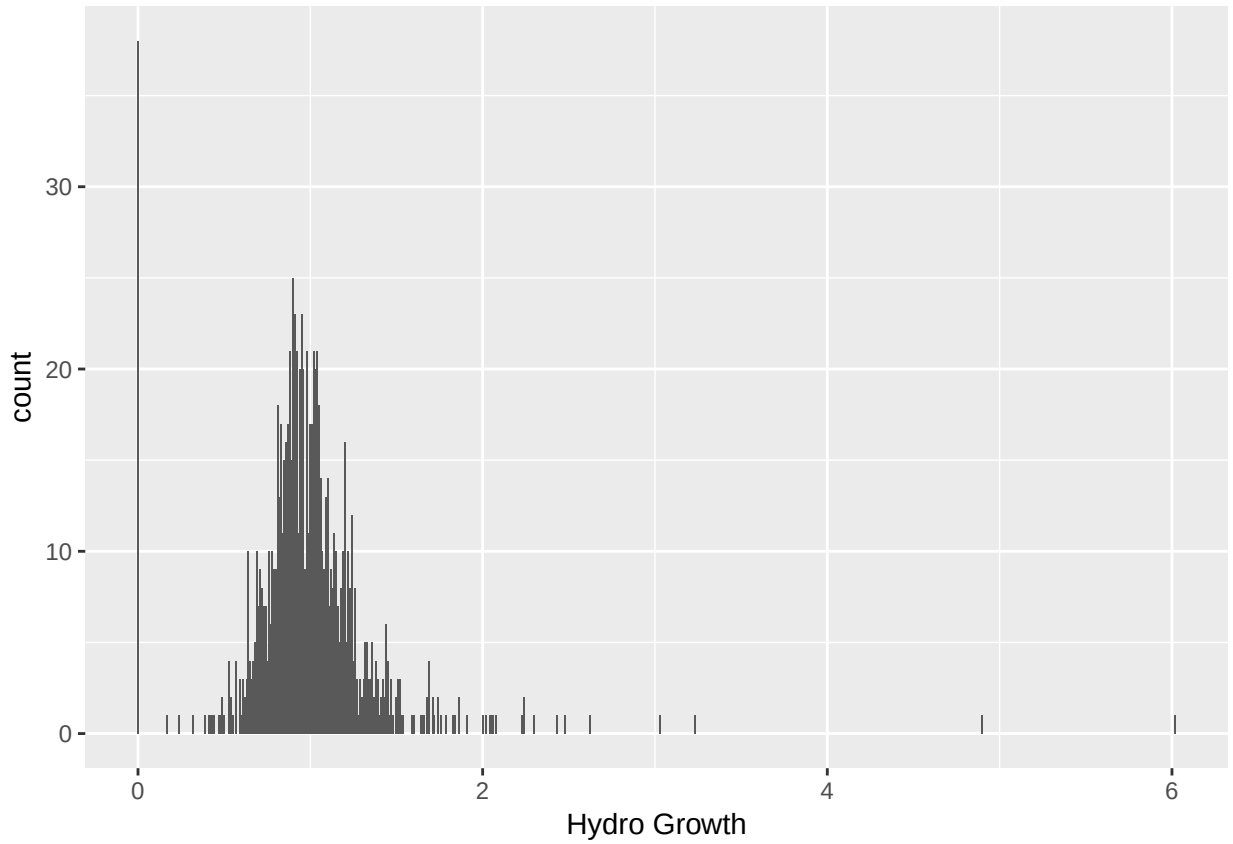

```
ggplot(df, aes(x=coal.growth)) +
  geom_bar()  + #looks normal (?)
  xlab("Coal Growth")
```



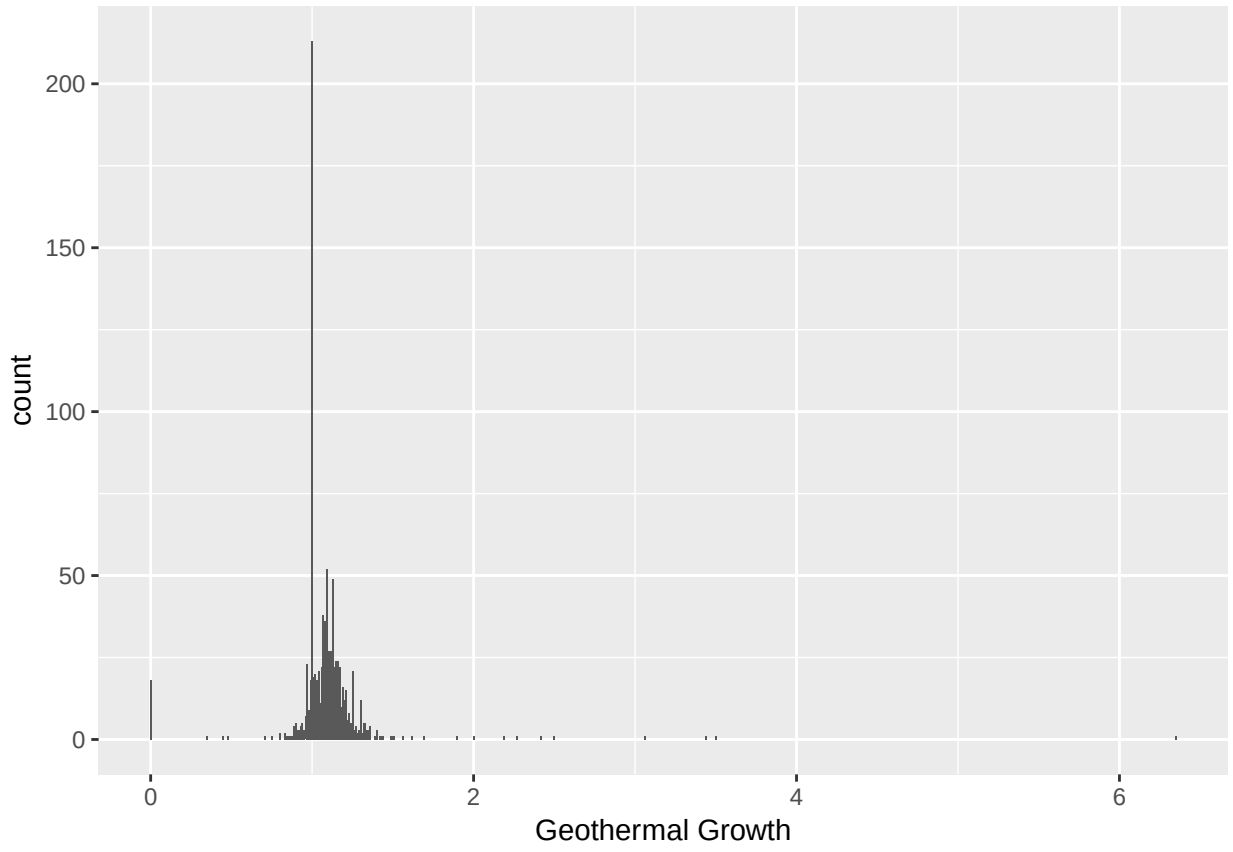

```
ggplot(df, aes(x=solar.growth)) +
  geom_bar() +
  xlab("Solar Growth")
```



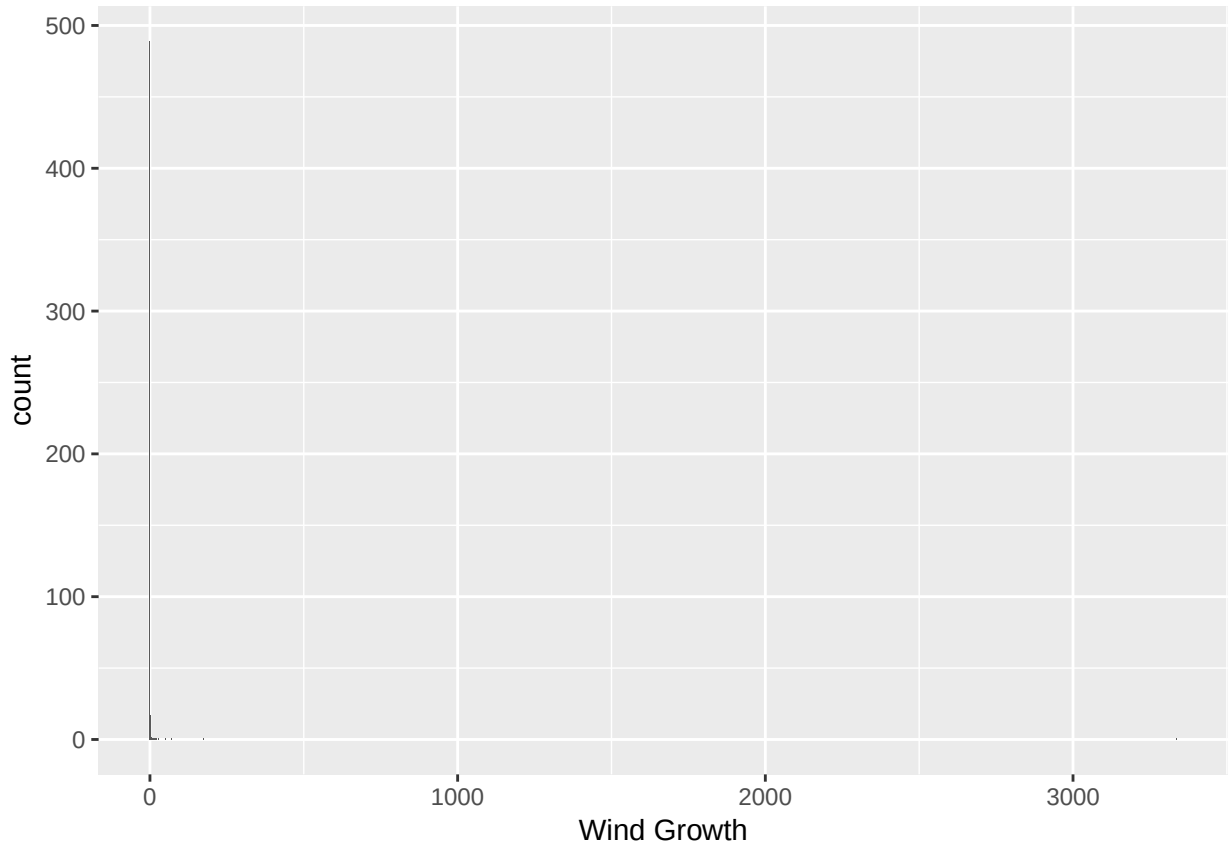

```
ggplot(df, aes(x=hydro.growth)) +
  geom_bar() +
  xlab("Hydro Growth")
```



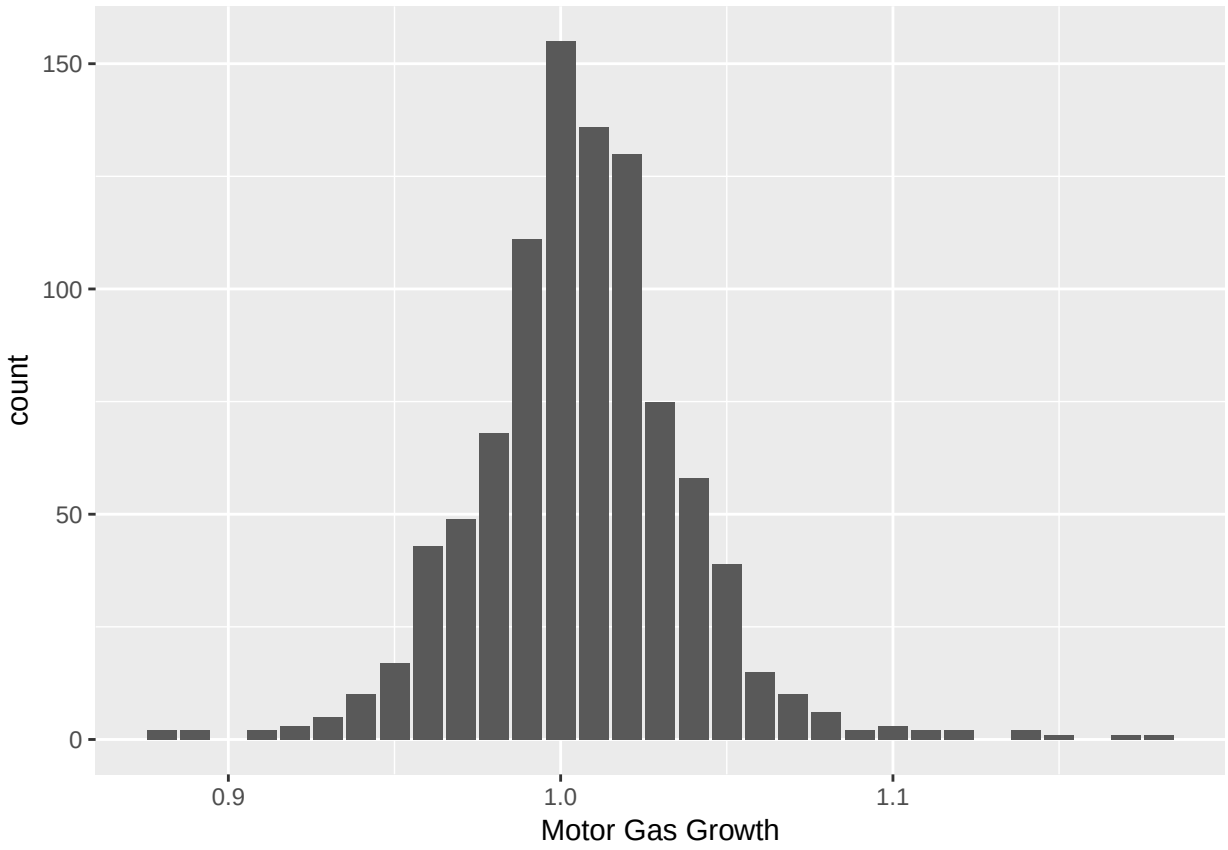

```
ggplot(df, aes(x=geothermal.growth)) +
  geom_bar() +
  xlab("Geothermal Growth")
```




```
ggplot(df, aes(x=wind.growth)) +
  geom_bar() +
  xlab("Wind Growth")
```

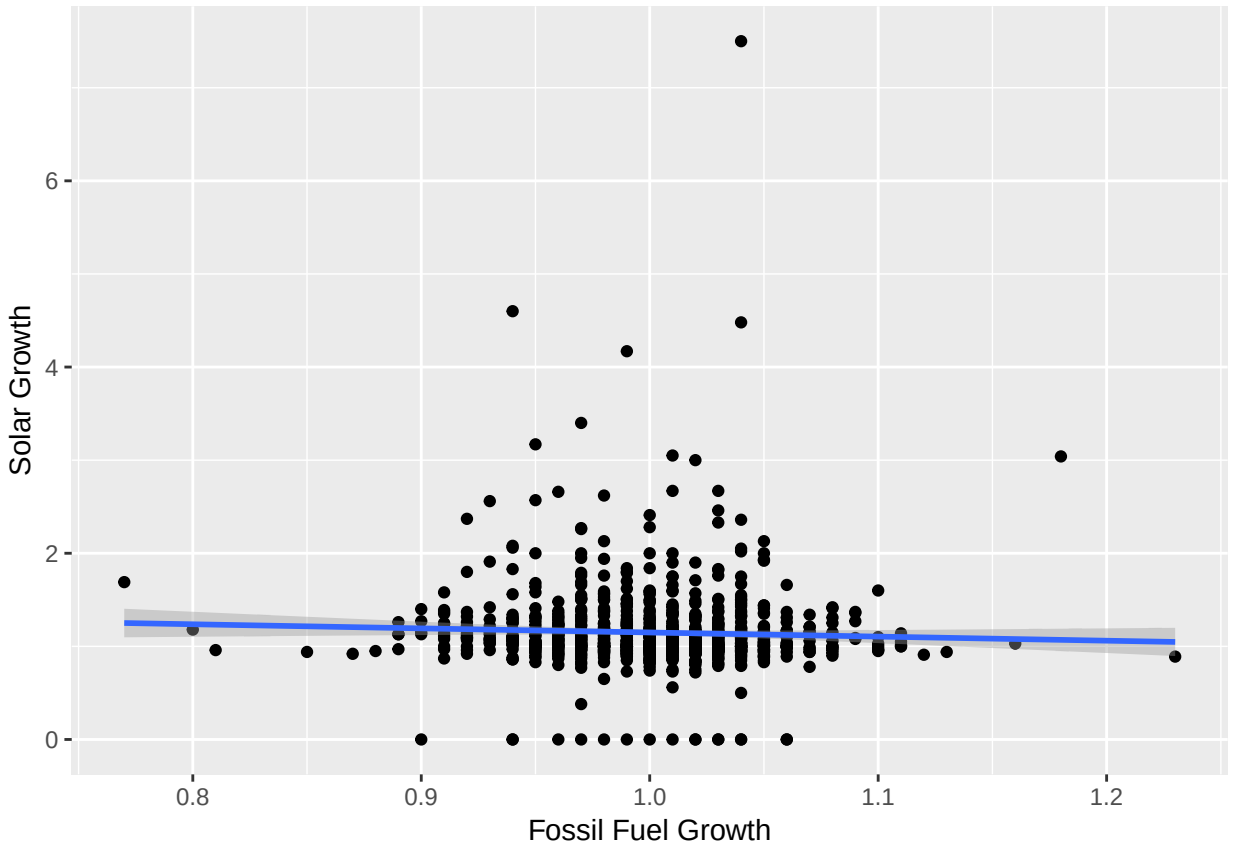



```
ggplot(df, aes(x=motor_gas.growth)) +
  geom_bar() +
  xlab("Motor Gas Growth")
```



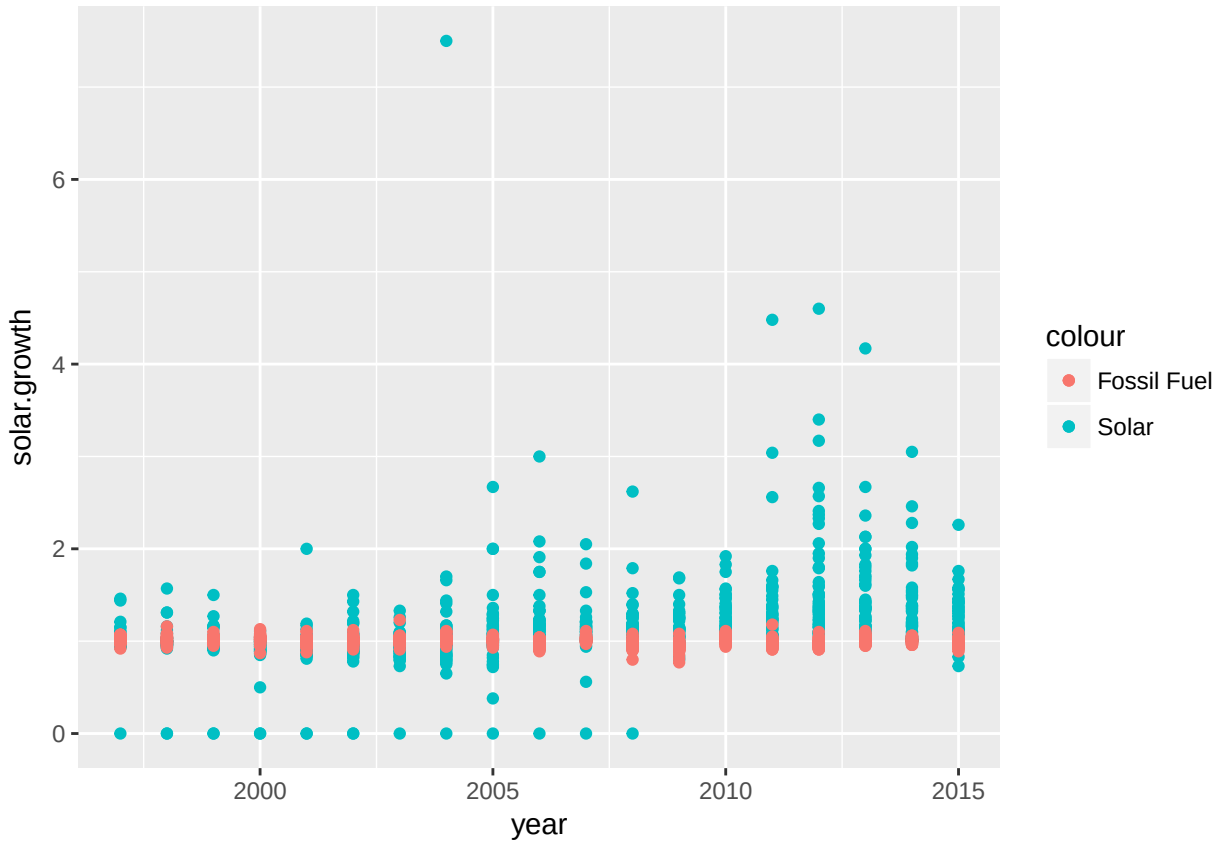
Hoping to see that when fossil fuel growth was less than 1, solar growth was above one. This is kind of the case, the regression line shows a decline of solar growth as fossil fuel increases, which makes sense. There are, however, a couple of points that do not follow this general rule.

```
ggplot(df, aes(x = fossil_fuel.growth, y = solar.growth)) +
  geom_point() +
  geom_smooth(method = lm) +
  xlab("Fossil Fuel Growth") +
  ylab("Solar Growth")
```

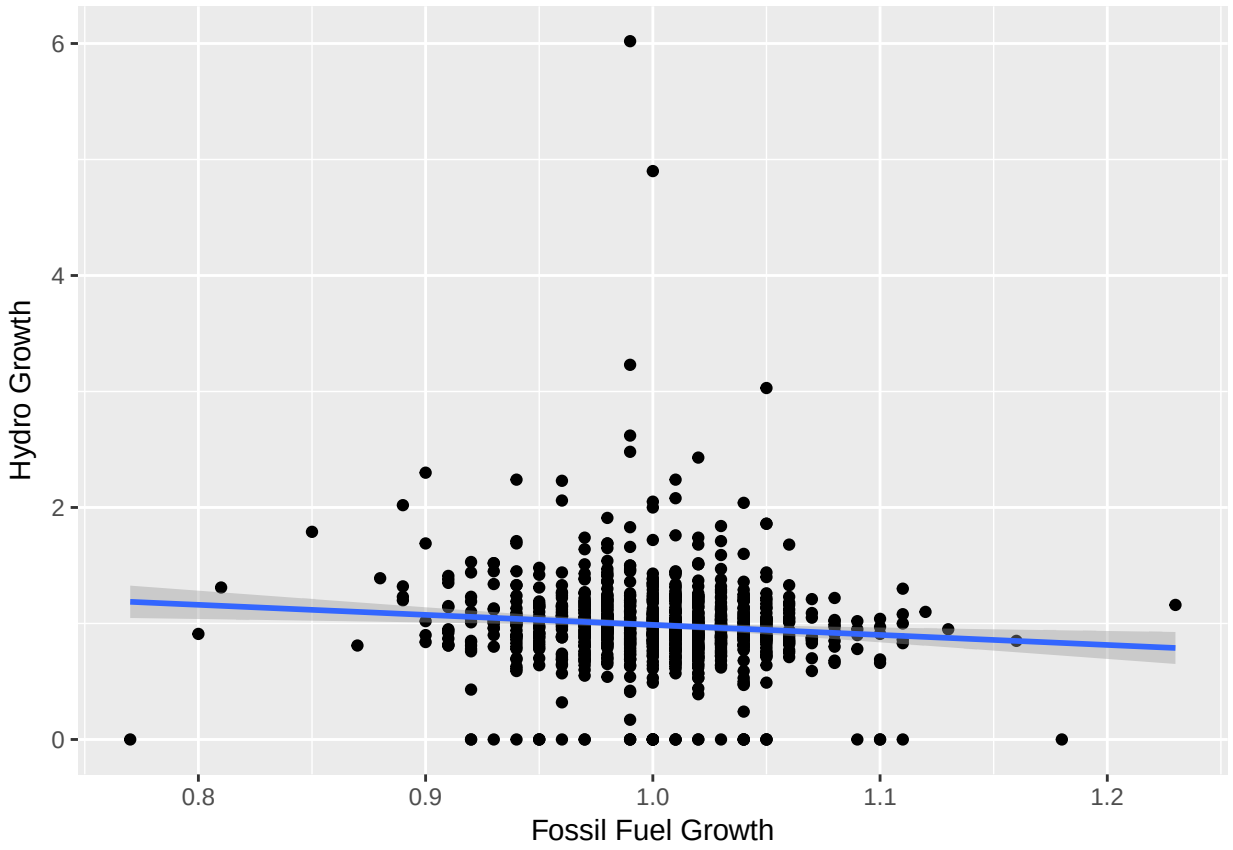


This graph is showing us an increase in solar energy growth and it does somewhat have a decreasing trend for fossil fuel growth as time goes one.

```
ggplot(df, aes(x = year, y = solar.growth)) +
  geom_point(aes(color = "Solar")) +
  geom_point(aes(y = fossil_fuel.growth, color = "Fossil Fuel"))
```

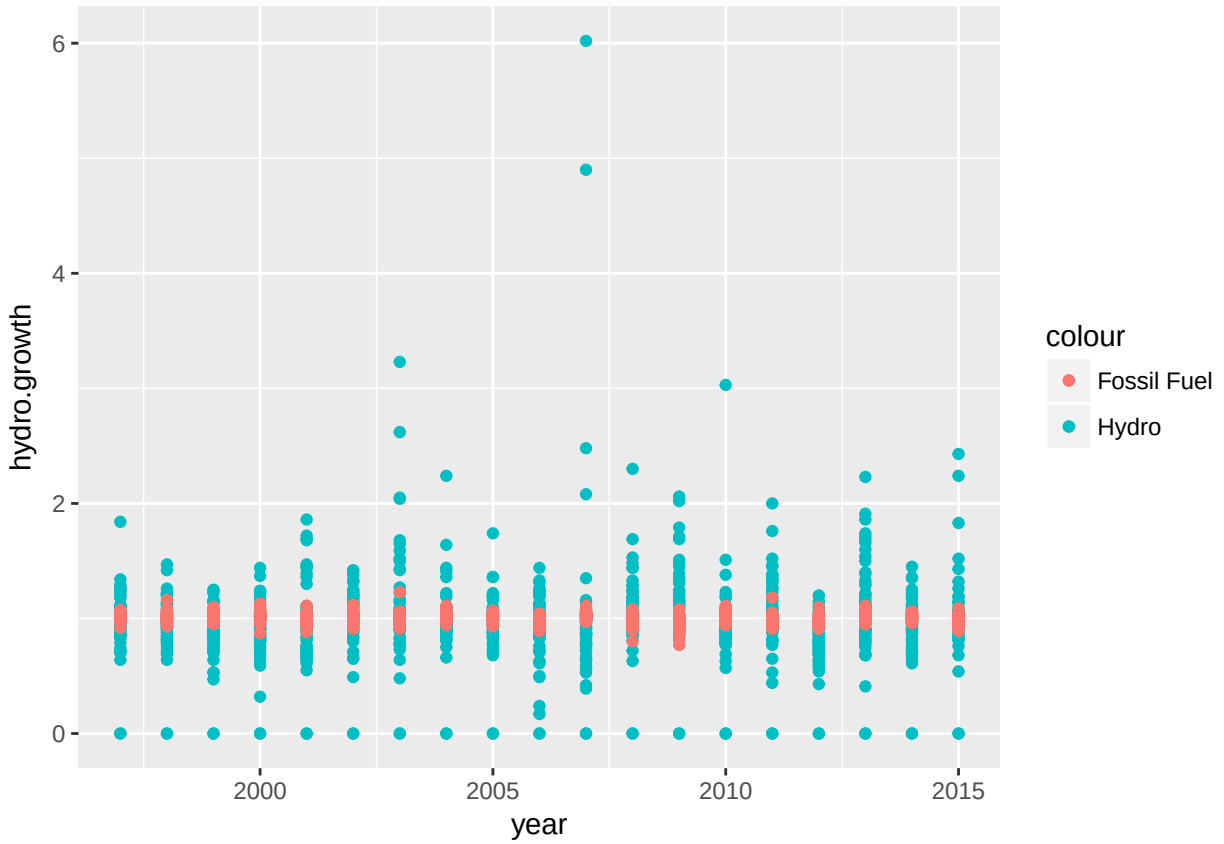



```
ggplot(df, aes(x = fossil_fuel.growth, y = hydro.growth)) +
  geom_point() +
  geom_smooth(method = lm) +
  xlab("Fossil Fuel Growth") +
  ylab("Hydro Growth")
```



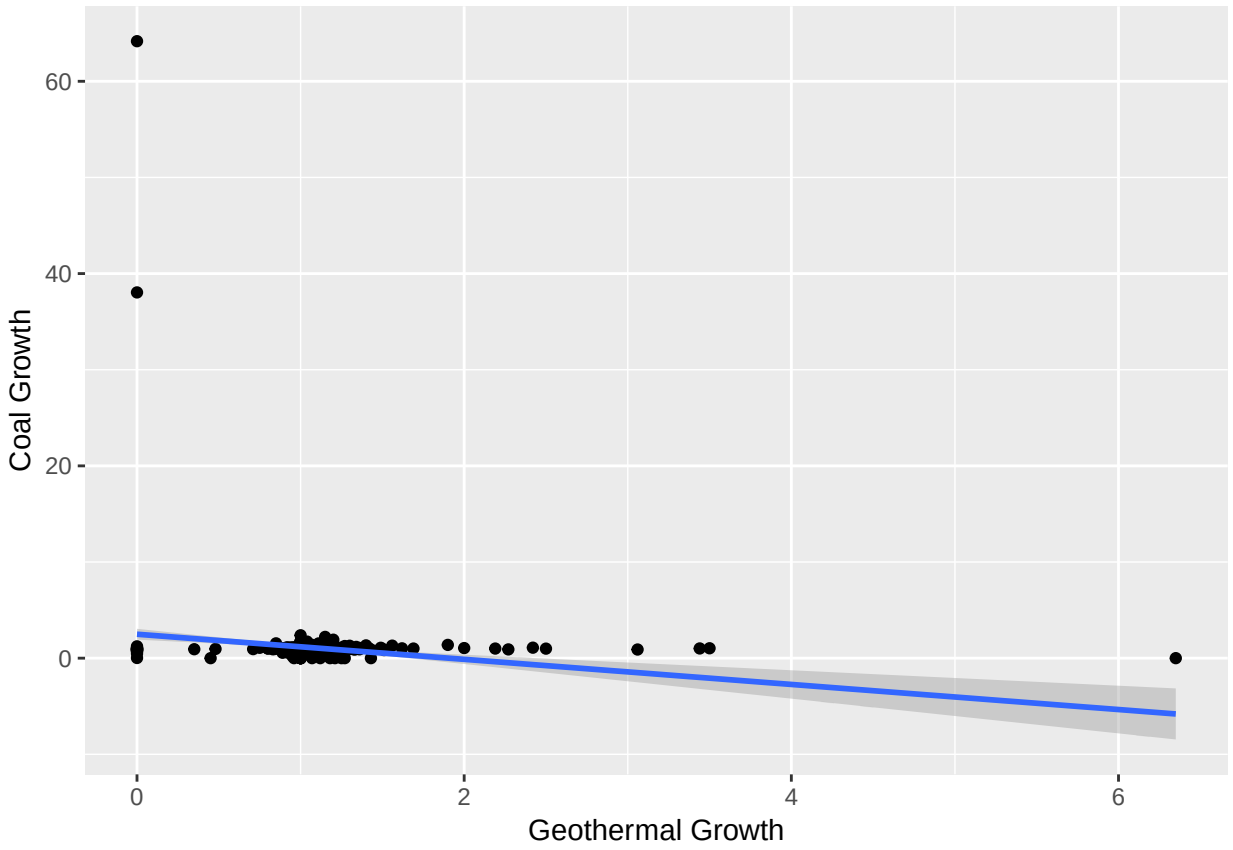
Hydro doesn't seem to be a replacement for the fossil fuel, as they seem to be continuously growing as time increases regardless of how the other is doing.

```
ggplot(df, aes(x = year, y = hydro.growth)) +
  geom_point(aes(color = "Hydro")) +
  geom_point(aes(y = fossil_fuel.growth, color = "Fossil Fuel"))
```



When coal growth was really high, the corresponding growth for geothermal was actually 0, which intuitively make sense because they would need to be producing a lot more if they don't have the choice to use another energy source).

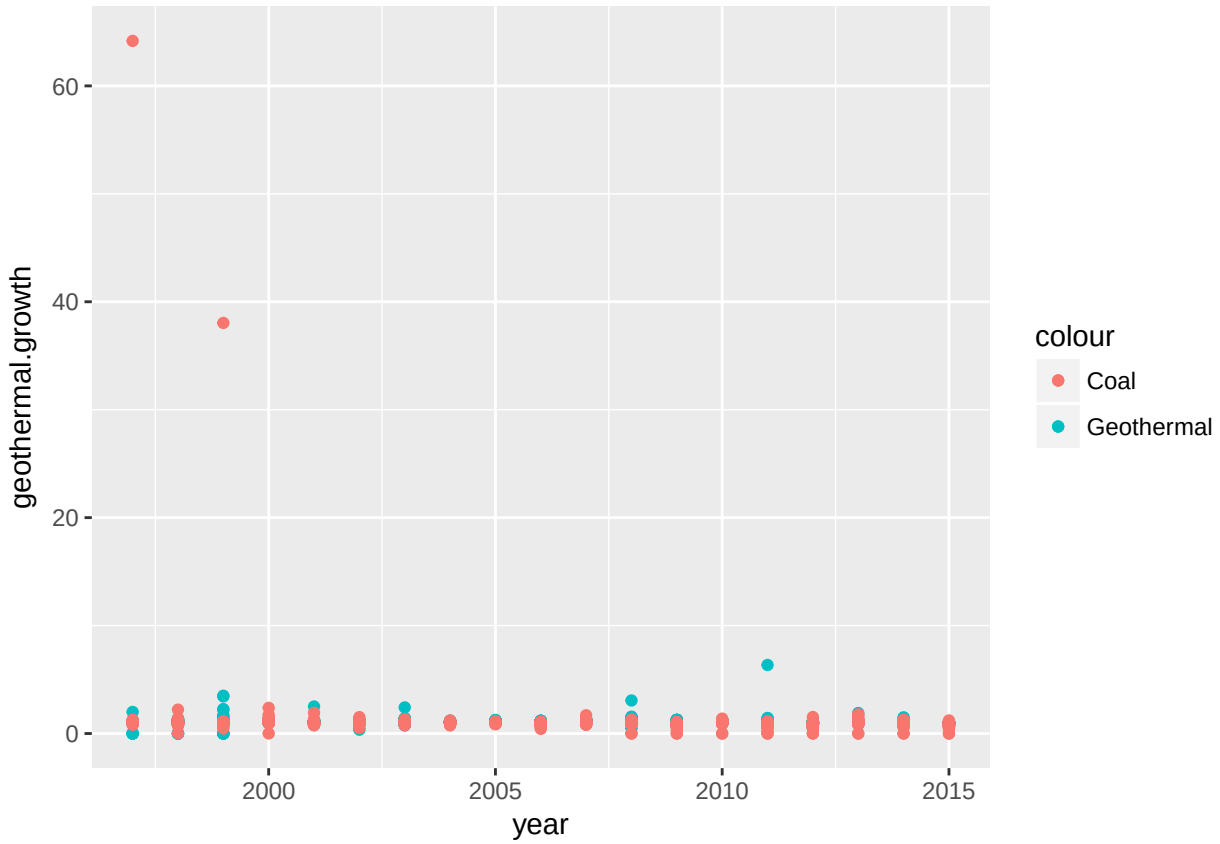
```
ggplot(df, aes(x = geothermal.growth, y = coal.growth)) +
  geom_point() +
  geom_smooth(method = lm) +
  xlab("Geothermal Growth") +
  ylab("Coal Growth")
```



From the previous graph we can see that geothermal was thriving and since it was above one for the geothermal growth, I personally think it's fine that coal growth was only around 1-ish (those outlines make it hard to see that).

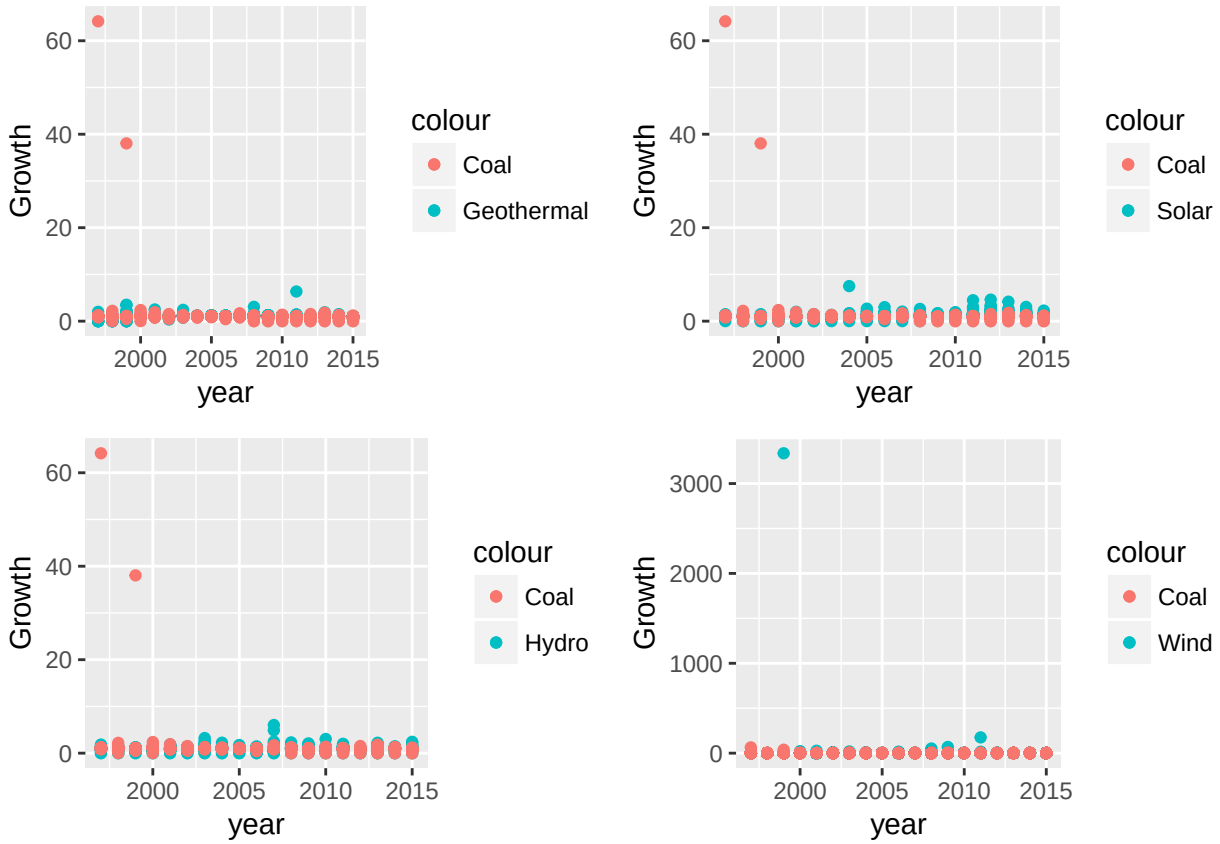
Coal has remained a steady growth throughout this time period, which makes me think that there hasn't been an implementation of another energy source that would change it. But also those outlines are making it very hard to see the trend.

```
ggplot(df, aes(x = year, y = geothermal.growth)) +
  geom_point(aes(color = "Geothermal")) +
  geom_point(aes(y = coal.growth, color = "Coal"))
```




```
#df$geothermal.growth[df$coal.growth >30] shows us that those two outliers have a corresponding geother
```

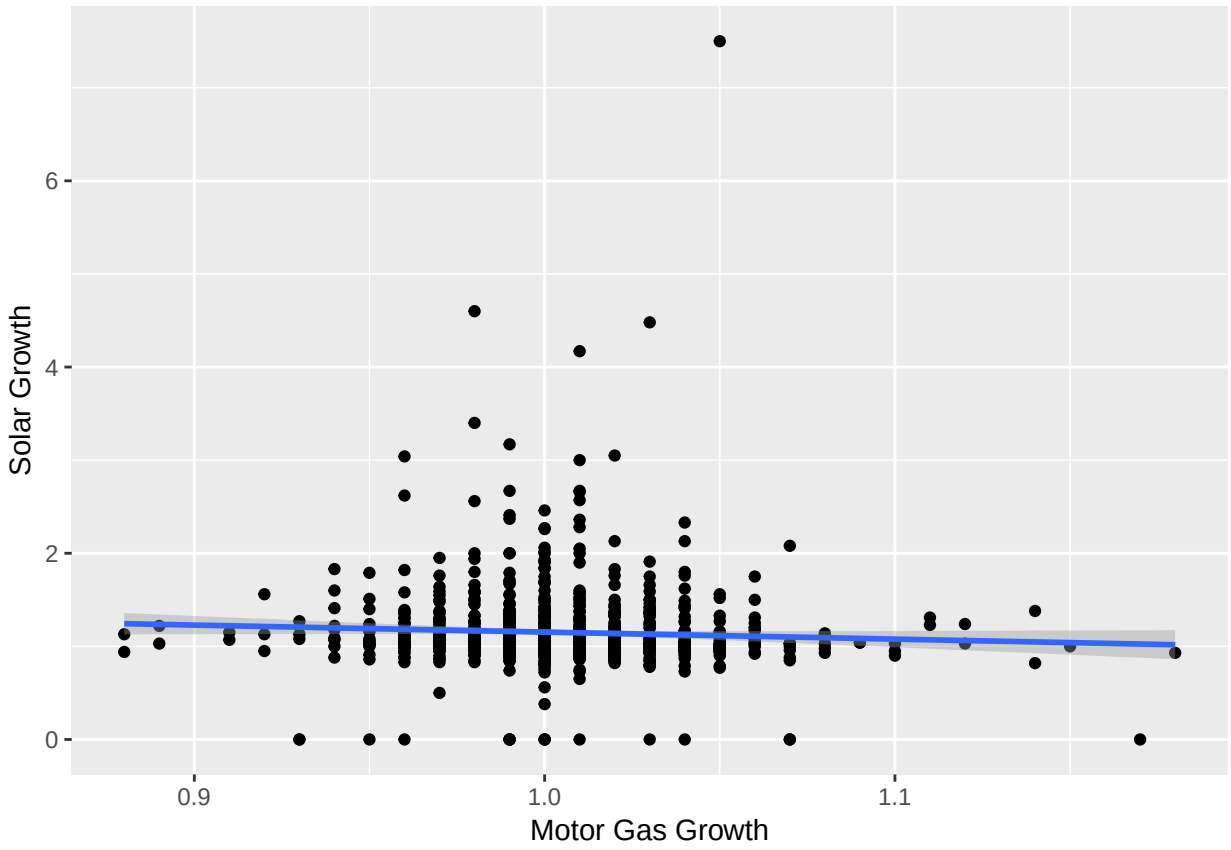
```
p1 <- ggplot(df, aes(x = year, y = geothermal.growth)) +
  geom_point(aes(color = "Geothermal")) +
  geom_point(aes(y = coal.growth, color = "Coal")) +
  ylab("Growth")
p2 <- ggplot(df, aes(x = year, y = solar.growth)) +
  geom_point(aes(color = "Solar")) +
  geom_point(aes(y = coal.growth, color = "Coal"))  +
  ylab("Growth")
p3 <- ggplot(df, aes(x = year, y = hydro.growth)) +
  geom_point(aes(color = "Hydro")) +
  geom_point(aes(y = coal.growth, color = "Coal")) +
  ylab("Growth")
p4 <- ggplot(df, aes(x = year, y = wind.growth)) +
  geom_point(aes(color = "Wind")) +
  geom_point(aes(y = coal.growth, color = "Coal")) +
  ylab("Growth")
grid.arrange(p1, p2, p3, p4, ncol=2)
```



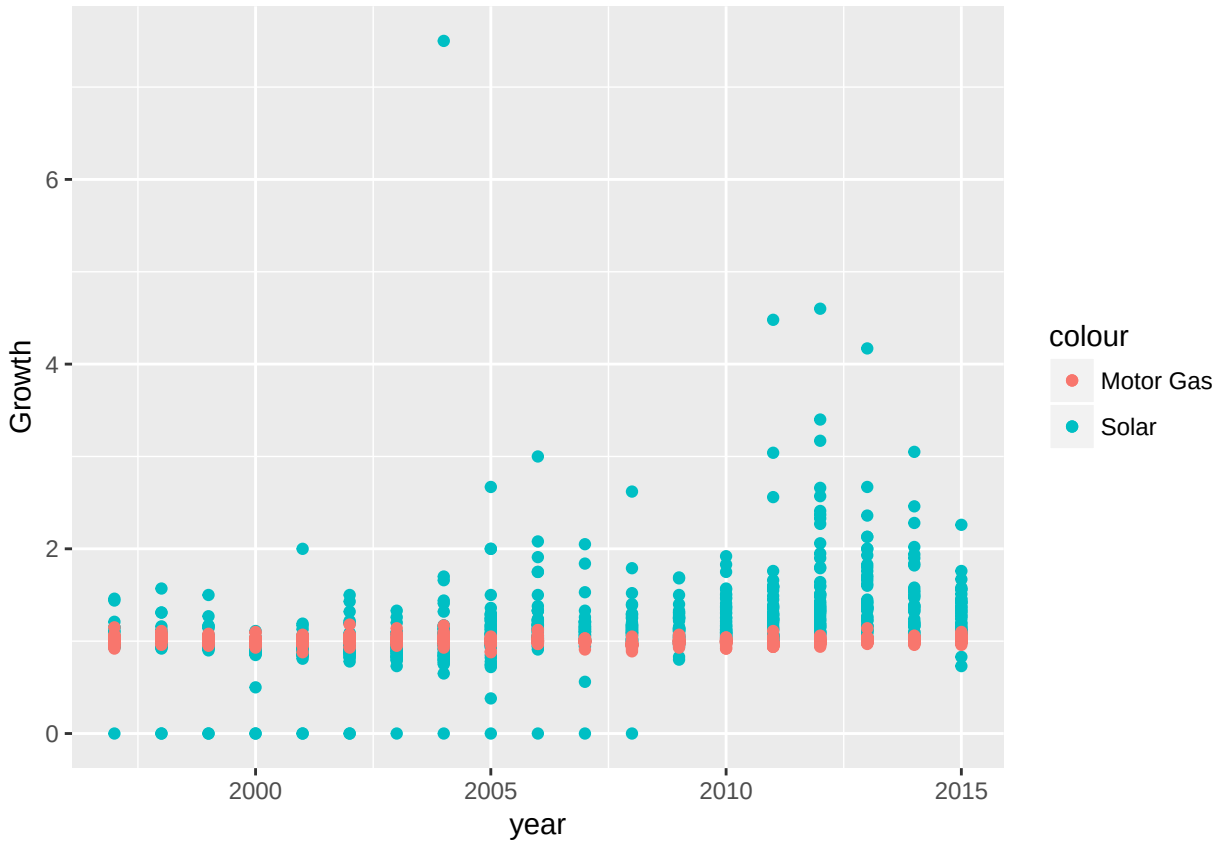
Wind is probably going to ruin every type of graph I create because of that outlier and so the rest of the graphs would make it hard to interpret.

We kind of see a decrease in the amount of growth in motor oil.

```
ggplot(df, aes(x = motor_gas.growth, y = solar.growth)) +
  geom_point() +
  geom_smooth(method = lm) +
  xlab("Motor Gas Growth") +
  ylab("Solar Growth")
```




```
ggplot(df, aes(x = year, y = solar.growth)) +
  geom_point(aes(color = "Solar")) +
  geom_point(aes(y = motor_gas.growth, color = "Motor Gas")) +
  ylab("Growth")
```


Growth
6
4
2
0
2000
2005
2010
2015
year
colour
Motor Gas
Solar